Members' experiences and questions about testing and assessment

How to make a judging plan for rated tests?

Jeffrey Durand

Testing situation

A few years ago, I had to put together a speaking test for all the students (about 2,000) at my university. About 60 teachers were available to rate students, who were tested in groups of four. Two teachers worked together to rate all the students in each group. In speaking tests, the raters are often not equally strict (some tend to give slightly higher scores than others), and on occasion may give an unusually high or low score. These problems can be discovered by using software like *Facets* (Linacre, 2012), and scores can be adjusted or students can be retested. To do this, however, there needs to be a way to know how strict each rater is in comparison to others. This can only be done if all the raters (and tasks and prompts) are connected together in what is called a judging plan (Linacre, 1997; Sick, 2013).

I found a pretty good judging plan while observing a colleague's speaking class. The instructor put students into two concentric circles, with equal numbers of students in each circle. A student in the outer circle worked with a partner from the inner circle. After a period of time, the students in the outer circle all rotated one place around the circle to talk with the next student in the inner circle. This created a regular ring lattice in which each student could be connected to all the others. Figure 1 shows a regular ring lattice with 16 raters (the blue diamonds), each with three partners (connected by straight lines). A slightly larger version of this method seemed to provide exactly what I needed for the raters. It also fit the testing location, which took place on two floors of a building that has stairwells at each end. The raters could quickly and easily move between rooms. After the judging plan was set, it was easy to randomly assign students to each room at a certain time.

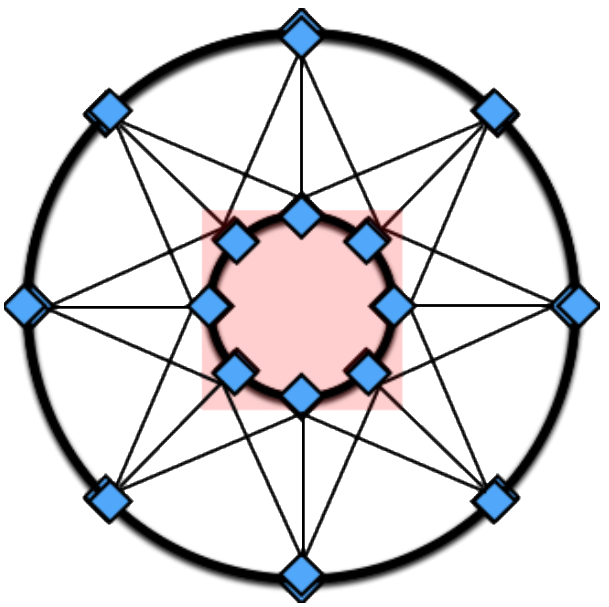

Figure 1. Judging plan

Questions

I have three questions about this judging plan.

1. How often should raters rotate, or for how many sessions should raters work together? Is it better to have raters work together for many sessions so that we are more confident about how strict they are in comparison to each other? Or should raters be rotated more often so that there are direct comparisons of strictness with more raters? Given that there is (thankfully) a limit to how many students an instructor is asked to rate, is there an optimal balance between rotating frequently and working with the same partner for a number of sessions?
2. Are there any other (better) ways of making a judging plan? For example, are there advantages of using three raters for each session or having an independent, trusted rater join random sessions? In your experience, what have been good (or not so good) ways of making judging plans?
3. Are there any questions that I have not considered that might be equally or even more important?

Do you have any real-life experience with judging plans or tests in which students are rated? Please share what you can!

References

Linacre, J. M. (1997). Judging plans and Facets. *MESA Research Note #3*, retrieved May 30, 2014, from http://www.rasch.org/rn3.htm.

Linacre, J. M. (2012). Facets (Version 3.70) [Computer Software]. Chicago: Winsteps.com.

Sick, J. (2013). Rasch measurement in language education part 7: Judging plans and disjoint subsets. *Shiken Research Bulletin, 17,* 27-31.

Where to Submit Questions:

Please send your responses to this question, as well as details about your own tests, to: tevalpublications@gmail.com

This section is a place for you, our readers, to share your experience with tests and to ask each other for advice. What you have learned can be a great help to others, both in the answers that you share and in the questions that you ask. When you submit your own questions about a test, remember to include a little background about it.